

# Analyzing Data for Systems Biology: Working at the Intersection of Thermodynamics and Data Analytics

William R. Cannon<sup>1</sup> and Douglas J. Baxter<sup>2</sup>

<sup>1</sup>Computational Biology and Bioinformatics Group, Fundamental and Computational Sciences Directorate

<sup>2</sup>Molecular Sciences Computing Facility, Environmental Molecular Sciences Laboratory,

Pacific Northwest National Laboratory, Richland, WA

Email: [william.cannon@pnnl.gov](mailto:william.cannon@pnnl.gov)

## Abstract

Many challenges in systems biology have to do with analyzing data within the framework of molecular phenomena and cellular pathways. How does this relate to thermodynamics that we know govern the behavior of molecules? Making progress in relating data analysis to thermodynamics is essential in systems biology if we are to build predictive models that enable the field of synthetic biology. We discuss work at the crossroads of thermodynamics and data analysis and demonstrate that statistical mechanical free energy is a multinomial log likelihood. Applications to systems biology are presented.

## 1. Introduction

Many challenges in systems biology have to do with analyzing data within the framework of molecular phenomena and cellular pathways. Many analogies have been used to describe the cell as a system, including electrical circuits [1], chemical factories [2], and computers [3]. While all these analogies have merit when evaluated in context, over 100 years of statistical physics has taught us that thermodynamics govern the behavior of molecules. There is no reason to believe that the laws of physics have been suspended in the case of biological phenomena.

How does data analysis of biological systems relate to thermodynamics that we know govern the behavior of molecules? Making progress in relating data analysis to thermodynamics is essential in systems biology if we are to build predictive models that enable the field of synthetic biology [4, 5].

Thus, thermodynamics should be a natural way to integrate the analysis of data with scientific models of cellular function. However, data analysis methods and protocols rarely use the language of statistical physics. Instead, bioinformaticists most often use the language and methods of statistics to describe and analyze biological data. Indeed, principled statistical approaches have led to clear and demonstrably better analysis methods than ad hoc procedures. Nevertheless, integration of statistical thermodynamics into the data analysis should in principle open the door for integrating scientific models tightly with analysis of large datasets typical of systems biology.

We discuss in this paper work at the crossroads of thermodynamics and data analysis and demonstrate that free energy is a statistical multinomial log likelihood. Applications to systems biology are discussed. As an example, the application of a thermodynamically

inspired log likelihood analysis to proteomics data analysis increases the number of spectra that can be identified and associated with biological processes by 50–150%.

## 2. Free Energy is a Multinomial Log Likelihood

Free energies are formulated from partition functions, while data analysis methods use probabilistic approaches such as likelihoods. This section examines each perspective and explores the different terminologies used in these approaches and the overlap between likelihoods and free energies. The relation between free energy, likelihood, and Shannon's entropy are examined.

*Data Analytics Perspective: Multinomial likelihoods.* First, we consider the analysis of data from a purely descriptive and statistical perspective. For the sake of demonstration, we consider assessing the likelihood of the distribution of products from a reaction,  $b \rightarrow a_1, \dots, a_k$ . In this example,  $n_{tot}$  identical reactant molecules,  $b$ , give rise to  $k$  different reaction products,  $a_1, \dots, a_k$ . In the reaction, each  $b$  molecule reacts to produce a product  $a_i$ , and  $\mathbf{n} = \{n_i\}$  is the vector count of the number of reactants,  $n_i \leq n_{tot}$ . The probability of forming product  $a_i$  depends on the molecular energy ( $\epsilon$ ) required for the chemical reaction and the temperature of the system; we denote this probability as  $\theta_i(\epsilon, T)$ . Each of the  $n_i$  products is independent and indistinguishable, but each of the products  $a_i$  is distinguishable from the other products. For product, the probability  $\theta_i$  and the data  $n_i$  can be used to form a multinomial model of producing the products

$$p(\mathbf{n} | \theta) = n_{tot}! \prod_j^k \frac{1}{n_j!} \theta_j(\epsilon, T)^{n_j} . \quad (1)$$

The term on the left-hand side is the likelihood,  $p(\mathbf{n} | \theta)$ , for observing the products distributed according to  $\mathbf{n}$  given the model parameters  $\theta$ . There are no terms  $(1 - \theta_i(\epsilon, T))$  when product  $a_i$  is missing in the equation above because these are accounted for by the probabilities for other products:  $(1 - \theta_i(\epsilon, T)) = \sum_{j \neq i} \theta_j(\epsilon, T)$ . Assuming that one can accurately

count the products using an experimental apparatus, the calculation of this distribution for any set of products is relatively straightforward in principle. However, this formulation is difficult to use in practice because the parameters  $\theta_i$  refer to the microscopic state and are not known. In fact, in most experimental measurements the energy levels of the individual molecules are not observed; instead, the average energy of a collection of molecules is measured. However, the analysis can instead be treated on the macroscopic level such that

$$p(\mathbf{n} | \theta) = n_{tot}! \prod_j^k \frac{1}{n_j!} \theta_j(T)^{n_j} . \quad (2)$$

In this case, the probabilities  $\theta_j(T)$  are measured as a function of temperature only, and energy levels are not resolved. The  $\theta_j(T)$  can be estimated from counts such that the model parameters for any product  $a_i$  are the number of observed products  $n_i$  out of  $n_{tot}$  fragments,

$$\theta_i(T) = \frac{n_i}{n_{tot}}.$$

That is,  $\theta_i(T)$  is the estimated probability of observing  $n_i$  molecules of product  $a_i$  out of  $n_{tot}$  product molecules when the measurement is done at temperature  $T$ .

*Thermodynamic Perspective: Partition Functions.* Statistical mechanics is the foundation for thermodynamics, which consists of the statistical description of the physics of the molecular processes and interactions. The key concept in deriving thermodynamic properties from energetics is the partition function,  $q_i$ , which accounts for energy levels  $l$  for each chemical species  $a_i$ . This function is given by the following.

$$q_i = \sum_l e^{-\varepsilon_{il}/RT}$$

The Helmholtz free energy can then be expressed in terms of partition functions.

$$-a/k_B T = \log \left( n_{tot}! \prod_j^k \frac{1}{n_j!} q_j^{n_j} \right)$$

Here,  $k_B$  is Boltzmann's constant,  $T$  is the absolute temperature in Kelvin, and  $q_j$  is the molecular partition function that accounts for energy levels for each product  $j$ . As in the data analysis description previously,  $n_{tot} = \sum n_j$ , and the vector quantities  $n = \{n_i\}$  are fixed. For convenience, we rewrite the free energy as

$$\begin{aligned} -a/k_B T &= \log \left( n_{tot}! \prod_j^k \frac{1}{n_j!} q_j^{n_j} \right) + \log \frac{1}{q^{n_{tot}}} - \log \frac{1}{q^{n_{tot}}} \\ &= \log \left( \frac{n_{tot}!}{q^{n_{tot}}} \prod_j^k \frac{1}{n_j!} q_j^{n_j} \right) - \log \frac{1}{q^{n_{tot}}}. \end{aligned} \quad (3)$$

The likelihood formulation used in data analysis can be recovered from this latter formulation. First, we note that the microscopic probabilities  $\theta_i(\varepsilon, T)$  used in the likelihood formulation are the Boltzmann probabilities,

$$\theta_i(\varepsilon_{il}, T) = \frac{e^{-\varepsilon_{il}/RT}}{\sum_j^k \sum_l^\infty e^{-\varepsilon_{jl}/RT}},$$

in which  $\varepsilon_{il}$  is the  $l$ th energy level for product  $a_i$  [6]. The denominator is the multiproduct partition function that accounts for all products  $j$ ,  $q = \sum_j^k \sum_l^\infty e^{-\varepsilon_{jl}/RT}$ . The partition function can

be recognized as the cumulative distribution of the likelihood of each product  $a_j$ , which measures the extent of the available state space. Marginalizing this probability over the available energy levels gives, for a given product,

$$\theta_i(T) = \frac{\sum e^{-\varepsilon_{il}/RT}}{q},$$

$$= \frac{q_i}{q}$$

where  $q_i$  is the molecular partition function discussed above. As a result, the likelihood of Equation 2 is proportional to the product of the molecular partition functions defined above,

$$p(\mathbf{n} | \theta) = n_{tot}! \prod_j \frac{1}{n_j!} \theta_j(T)^{n_j}$$

$$= n_{tot}! \prod_j \frac{1}{n_j!} \cdot \left( \frac{q_j}{q} \right)^{n_j}$$

$$= \frac{n_{tot}!}{q^{n_{tot}}} \prod_j \frac{1}{n_j!} \cdot q_j^{n_j}$$

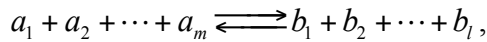
Substituting this equality into the left-hand side of Equation 3 gives

$$-a/k_B T = \log \left( \frac{n_{tot}!}{q^{n_{tot}}} \prod_j \frac{1}{n_j!} q_j^{n_j} \right) - \log \frac{1}{q^{n_{tot}}}$$

$$= \log p(\mathbf{n}, V, T | \theta) + n_{tot} \log q$$

This says that the free energy of a system with  $\mathbf{n} = \{n_i\}$  species, volume  $V$ , and temperature  $T$  depends on the likelihood  $p(\mathbf{n}, V, T | \theta)$  and the total state space, or likelihood distribution, available to the molecules. The only assumption about equilibrium regarding  $\theta_i$  is that the degrees of freedom in the physical system are coupled such that energy can be transferred among the species; that is, it is assumed that the  $\theta_i$  reflect the true Boltzmann probabilities. The second term on the right-hand side elucidates the primary difference between free energy and a data analysis approach, and that is a difference in normalization. While the statistical likelihood uses the cumulative distribution among the products as a normalization factor, free energies are based on a physical scale referenced to a temperature of absolute zero and perfect order.

For a general chemical reaction given by



the relative free energy that determines whether the products or reactants will be observed is given by [7]

$$\Delta a / k_B T = -\log \frac{p_b(n_b, V, T | \theta)}{p_a(n_a, V, T | \theta)} - \log \frac{q_b^{n_{b,tot}}}{q_a^{n_{a,tot}}} \quad (4)$$

While the first term on the right-hand side compares the likelihood of each served distribution, the second term compares the extent of each state space including the distance

of each distribution from perfect order at absolute zero. The first term on the right hand side is the likelihood ratio often used in data analysis:

$$LR = -\log \frac{p_b(n_b, V, T | \theta)}{p_a(n_a, V, T | \theta)}$$

$$= -\log \frac{n_{b,tot}! \prod_j \frac{1}{n_{b,j}!} \theta_{b,j}^{n_{b,j}}}{n_{a,tot}! \prod_j \frac{1}{n_{a,j}!} \theta_{a,j}^{n_{a,j}}}.$$

A typical goal in data analysis is to infer which of two scientific models of the observed phenomena best explains the data. In this scenario, the competing scientific models are represented by the parameters  $\theta_{a,j}$  and  $\theta_{b,j}$ , respectively, and the observations are represent by  $n_{obs} = n_a = n_b$ . The log likelihood ratio simplifies to

$$\log LR = \log \prod_j \left( \frac{\theta_{a,j}}{\theta_{b,j}} \right)^{n_j}$$

$$= \sum_j n_j \log \left( \frac{\theta_{a,j}}{\theta_{b,j}} \right).$$

Strictly speaking, the  $n_i$  refer to counts of molecules. However, using raw counts obtained from experimental measurements for  $n_i$  can be problematic if the values of  $n_i$  are large. For example, if the estimated counts are in the femtomolar range ( $10^8$  molecules), then this can lead to quite large values of the likelihood. The counts could be converted to molar values but with the opposite effect that the likelihood values become quite small. An intermediate solution is to scale the estimated counts by the total counts,  $n_i/n_{tot} = \rho_i$ . Scaling the counts in such a manner leads to the information theory entropy [8]

$$\frac{1}{n_{tot}} \sum_j n_j \log \left( \frac{\theta_{a,j}}{\theta_{b,j}} \right) = \sum_j \rho_j \log \left( \frac{\theta_{a,j}}{\theta_{b,j}} \right). \quad (5)$$

While Equation 5 can be interpreted within Shannon's framework [8] as the relative entropy between the probability space for model  $A$  of the observed phenomena and the probability space for model  $B$ , a more intuitive interpretation is that it represents the relative likelihood of comparable phenomena for the two scientific models, as averaged over the observed data.

One can obtain insight into the likelihood analysis by analogy with thermodynamic properties. For example, a "data" potential can be derived in analogy with the chemical potential ( $u_i = \frac{\partial A}{\partial n_i}$ ) from the likelihood ratio as follows.

$$\frac{\partial}{\partial \rho_i} \log LR = \frac{\partial}{\partial \rho_i} \sum_j \rho_j \log \left( \frac{\theta_{a,j}}{\theta_{b,j}} \right)$$

$$= \log \left( \frac{\theta_{a,i}}{\theta_{b,i}} \right)$$

This equation tells us that the influence of changing counts of product  $i$  on our average likelihood ratio is simply the log likelihood of the parameters for product  $i$  in the two models. Analogies to entropy ( $\frac{\partial A}{\partial T}$ ) and pressure ( $\frac{\partial A}{\partial V}$ ) are not directly possible; but temperature and volume are directly related to the Boltzmann probabilities, and we can examine the change in a model probability,

$$\begin{aligned}\frac{\partial}{\partial \theta_{a,i}} \log LR &= \frac{\partial}{\partial \theta_{a,i}} \sum_j^k \rho_j \log \left( \frac{\theta_{a,j}}{\theta_{b,j}} \right) \\ &= \frac{\rho_i}{\theta_{a,i}}\end{aligned}$$

That is, the influence on the log likelihood of changing the model probability for product  $i$  is directly proportional to the ratio of the observed counts and the probability for product  $i$ .

### 3. Applications to Systems Biology

Approaches such as this will ultimately have application to hypothesis testing of large data sets from systems biology. For example, computational models of metabolism, such as flux balance analyses and stochastic simulations, can often make predictions about which metabolites will be present. Competing models of the metabolism of a bacterial cell can then be compared. In this case, the model parameters  $\theta_{a,j}$  and  $\theta_{b,j}$  are the predicted abundances of metabolite  $j$  in the respective models, and  $n_j$  or  $\rho_j$  are the experimental observations from a metabolomics study. Using a test of significance, one can then reject one of the models.

As a demonstration, we have applied this approach to proteomics studies, in which the challenge is to select the peptide that best explains the experimental data. In this case, the experimental data are peak abundances from tandem mass spectrometry of unknown peptides. Each peptide present in the genome of the organism represents a competing scientific model in the form of a model spectrum. In each model spectrum,  $\theta_{a,j}$  represents the probability of observing fragment  $j$  of that peptide. Equation 5 is then used to decide which peptide is truly present, which in turn provides information on which protein is truly expressed under the conditions of the experiment.

Using this approach and highly accurate model spectra, one can obtain not only greater specificity but also greater sensitivity: the number of spectra that can be identified with a peptide is more than doubled compared with standard approaches [9]. The greater sensitivity and specificity are a direct result of the principled manner in which the model parameters and observed count information are evaluated. As mentioned above, each fragment  $i$  of a peptide is expected to be observed with a probability  $\theta_{a,i}$ , while  $n_i$  is the actual observed count. Each peptide is then evaluated as to whether it is a good match the observed spectrum by comparing the model spectrum of the peptide with a random (null) model spectrum.

$$LR = \sum_j^k \rho_j \log \left( \frac{\theta_{a,j}}{\theta_{0,j}} \right)$$

The null probability  $\theta_{0,i}$  of each peak  $i$  for the random model spectrum is derived by averaging probabilities for a peak at the same location from all other peptides that can be

derived from the organisms genome [10]. The principled evaluation of the count information and model probabilities used in scoring is important because model spectra for peptides vary tremendously in quality, and a scoring metric is needed that can maximally differentiate peptides based on their model spectra. For example, model spectra based on spectral libraries improve the identification rate by 50–150% compared with model spectra derived from statistical training over diverse sets of peptides [9].

Using this approach, we have been able to dramatically increase the number of spectra that can be matched to biological processes in *Synechococcus* sp. PCC 7002 [9], a cyanobacterium that is a model organism for studies of photosynthetic carbon fixation and biofuels development. Figure 1 depicts the coupled processes involved in photosynthesis and carbon fixation, of which many of the proteins are found in the accompanying table. Photosynthesis results in the splitting of water into protons and O<sub>2</sub>. Protons are also generated during oxygenic respiration, resulting in a proton gradient across the thylakoid membrane. This proton gradient is used to generate ATP, which is in turn used to fix CO<sub>2</sub> and synthesize larger reduced carbon species. Ultimately, six turns of the Calvin cycle result in one glucose molecule. The proteins from these cellular processes [11]—light harvesting for photosystem II (phycobilisomes), chlorophyll biosynthesis, CO<sub>2</sub> fixation (Calvin Benson Cycle), CO<sub>2</sub> uptake, and photorespiration—showed the large increases in number of identified peptides resulting from the use of the methods discussed above. The number of peptides identified within these five subsystems increased 50–60%. Furthermore, the number of peptides identified for photosystem I increased 160%, while the number of peptides associated with the pentose phosphate pathway increased 250%.

Also shown at the bottom of Figure 1 is a list of the individual proteins that showed the largest number of identified peptides. Notably, there was a 30,000% increase in the number of peptides identified for the CO<sub>2</sub> transporter of the ICT family. Bicarbonate and carbonate cycling is directly integrated with photosynthesis: CO<sub>2</sub> is the carbon source for the photosynthetically driven synthesis of sugars via the Calvin-Benson cycle. The key enzyme of the Calvin-Benson cycle is Ribulose-1,5-bisphosphate carboxylase oxygenase, or RuBisCO, which we were able to identify 30–60% more frequently. The reason for the increased number of identifications is directly linked to higher likelihood ratio scores due to the more realistic spectral model for the spectral libraries. The peptides showing the largest increases in matches to spectra were not necessarily high-abundance peptides or low-abundance peptides but, rather, peptides that were marginally identifiable using a standard database search. The observation that many of these peptides were involved in biological processes of interest, namely, carbon dioxide concentration and reduction, suggested that the methods discussed here will be important for increasing the accuracy and precision of proteomics to elucidate biological responses in general.

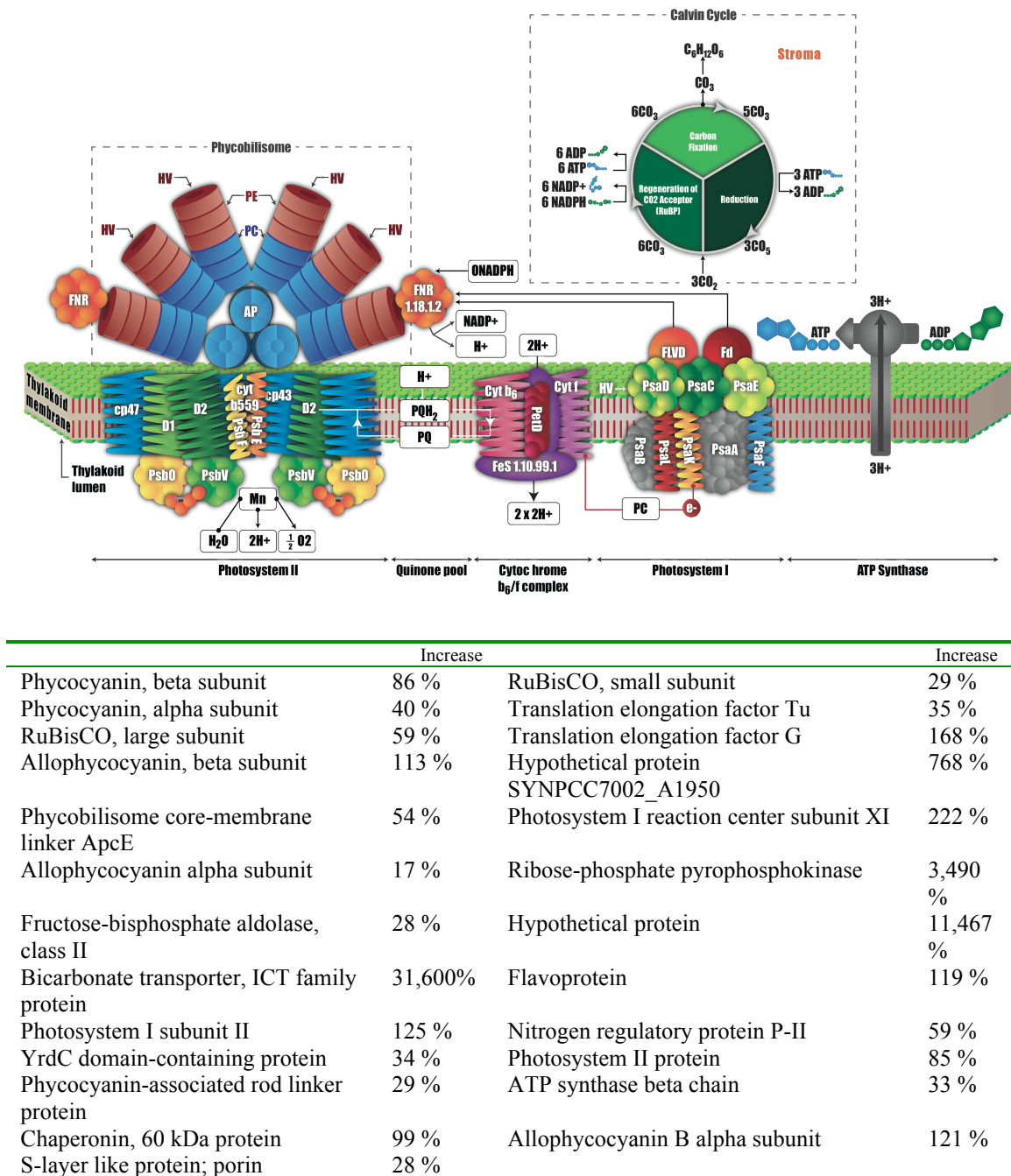


Figure 1. Use of a statistical log likelihood method to compare model spectra for peptides with experimental spectra for the cyanobacterium *Synechococcus sp.* PCC 7002, a model organism for biofuels development, led to a dramatic increase in the number of spectra that were identifiable with processes involved in photosynthesis and carbon fixation. The bottom panel lists the proteins that are identified with the most spectra, an indicator of protein abundance, and the percentage increase in identifiable spectra obtained when accurate model spectra are analyzed by using the thermodynamically inspired log likelihood.

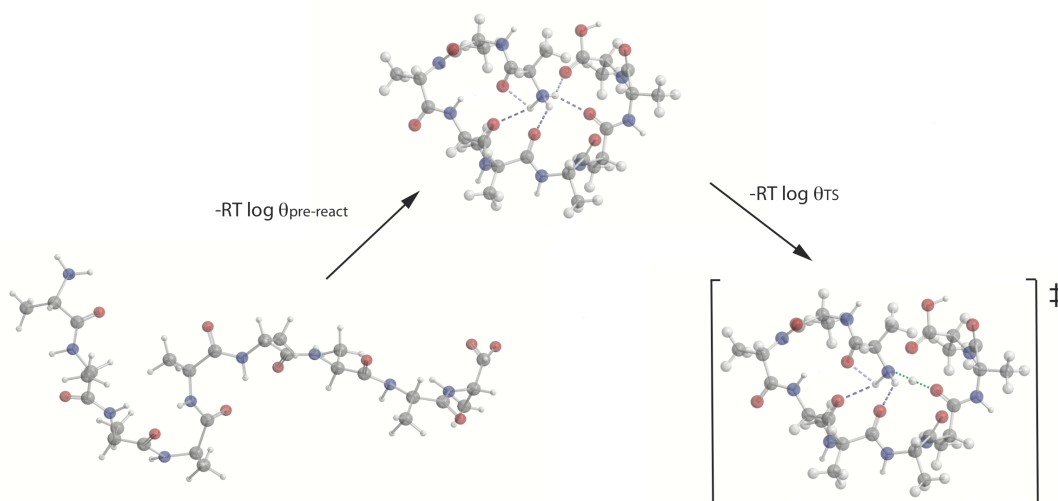


Figure 2. Structural reorganization involved in fragmentation at a peptide bond. Work must first be done to move the peptide from random configurations (left) into one in which the peptide is organized (center) for the bond-breaking steps (right). The free energy involved is proportional to the statistical likelihood of reaching each configuration.

*Informing Data Analysis with Physics-based Simulations.* Unfortunately, spectral libraries are not available for the majority of peptides. However, it may be possible to improve model spectra for data analysis by informing the models with physics-based simulations. In this case, physical models of the gas phase dynamics of the respective peptides are the relevant scientific models that can inform the data analysis [12]. The log likelihood scoring function outlined above allows for the use of probability distributions obtained from simulations. To demonstrate how these probabilities can be obtained, it is necessary to break down peptide fragmentation into two conceptual processes, shown in Figure 3, consistent with the mobile proton model of fragmentation [13]. The first process is the organization of the vibrationally excited gas phase peptide from random configurations into a configuration from which a proton can be but is not yet transferred from one of the proton donor groups (N-terminal amine or protonated side chain such as lysine or arginine) to a backbone carbonyl oxygen. This process can involve large-scale structural reorganization of both backbone and side chain groups. We will refer to this organized state primed for proton transfers as the pre-reactive state. The second process is the transfer of the proton to the carbonyl oxygen and the stretching and bending of bonds required to cross the transition state energy barrier and form products. Characterization of the transition state energies is exceedingly difficult even with the use of large compute resources because of the multitude of reaction channels that can lead to products. If one reaction channel dominates, then calculation of the transition state energy is much easier, but still a time-consuming and potentially labor intensive calculation. However, in the special case that fragmentation at each peptide bond occurs through reaction channels whose relative transition state energies do not change as a function of the position

of the fragmenting bond along the peptide backbone or the specific amino acids involved, then it may be reasonable to estimate an average likelihood of crossing the transition state  $\theta_i(T)_{TS}$  from the prereactive state that can be applied to all labile bonds. In this case, the likelihood of fragmenting a bond can then be estimated from combining the likelihood of reaching the pre-organized state with the likelihood of crossing the transition state.

$$\theta_i(T)_{fragmentation} = \theta_i(T)_{pre-react} * \theta_i(T)_{TS}$$

The likelihood of reaching the pre-reactive state from random configurations can be estimated from molecular simulations [7].

*Environmental Proteomics.* The development of highly accurate data analysis methods allows one to address areas have been previously challenging. One of these areas is the analysis of proteomics data from environmental samples. Organisms (microbes) in these samples may not have been previously identified; and even if they have been, they have typically never been sequenced. Yet matching model spectra of peptides to experimental spectra requires an a priori set of peptides, which are usually obtained from genome sequences.

Using the methods discussed above, we can now approach this problem by comparing the spectra of the environmental organisms with the genomes of all organisms that have been sequenced to date. Currently, there are just under 2,000 fully sequenced genomes representing a variety of organisms.

The approach that we have taken is to use optimization methods to match peptides and proteins from fully sequenced microbial genomes to the experimental spectra [14]. The method searches all fully sequenced genomes and optimizes proteome-spectra matches by iteratively eliminating microbes that are not likely to be in the sample. The method has been tested using samples containing blind mixtures of spectra from known microbes and samples containing unknown mixtures of microbes. The ability to analyze all fully sequenced genomes, however, requires analysis of up to 2,000 genomes which is roughly 6-10 million proteins and orders of magnitude more peptides against 10,000-to 30,000 or more spectra. This analysis requires high-performance computing.

#### 4. Programming Model and Scaling

*Programming Model.* The methods discussed above have been implemented for the analysis of proteomics data in serial, parallel [9], and map-reduce implementations [15]. In the serial code an input parameter file is read, along with the fragmentation model for generating model spectra, the protein sequences of all organisms to be searched, and the spectra to be analyzed. The code loops over the spectra, scoring each against all peptides generated from the protein sequences that are consistent with the observed mass-to-charge ratio of the intact peptide reported in each spectrum. The program accumulates high scoring matches and prints them out in a list, sorted by likelihood ratio score. The amount of work required to score a set of spectra depends on a variety of factors that include the number of candidate peptides to be analyzed per spectrum, the length of the peptides, the number of peaks from each peptide that match peaks from the experimental spectrum, and the number of spectra to analyze. The time for analysis of each spectrum cannot be predetermined without doing two-thirds of the work

required to actually score the spectrum, which makes an a priori determination of run time for a spectrum impractical. Hence, we use a dynamic scheduling scheme facilitated by a server/client process model.

The parallel version of the *MSPolygraph* is essentially a task-scheduling wrapper around the serial version of the code. In the parallel version we use the MPI (Message Passing Interface) standard for communication. The input files are placed on a globally visible file system (mounted on all the compute nodes). Each processor reads in to its own local memory the input files, and we then use a dynamically scheduled server-client model to control which process (mpi rank) scores which spectrum. A processor's behavior is controlled by its mpi rank. One processor is the dedicated server process (mpi rank 0), and all others are considered client processors.

After reading the input data, the server process issues a nonblocking receive to each client. It uses a simple counter to determine which task to send to a requesting client. It polls clients for responses indicating that a spectrum has been completed (or during the first pass that a client has initialized and is ready to start scoring) and replies with another index for a spectrum to be scored if there is one or a quit message if all tasks have been distributed. The manager utilizes nonblocking sends so as not to need to wait for the clients to receive their messages. A different buffer for each client is used. Since the outgoing messages are only an index as to which spectra to score, and the incoming messages are a fixed length summary line, only a small amount of space is required even for a very large number of clients. Also, since the server hands out a new index to a “ready to start” or “completed spectrum” message from a client, no more than one message per client is ever in flight. After all tasks have been handed out, the server processor continues polling for responses till responses for all spectra have been received.

Each client process is initiated by reading the input files from the global file system, and opening its own output file for printing results. It then issues a nonblocking receive for an index of the next spectra to be score and sends a “ready to start” message to the server processor. The client then enters a communication and processing loop in which the client repeatedly does the following:

1. Waits for server message indicating which spectrum to process (or a quit message).
2. Processes the spectrum

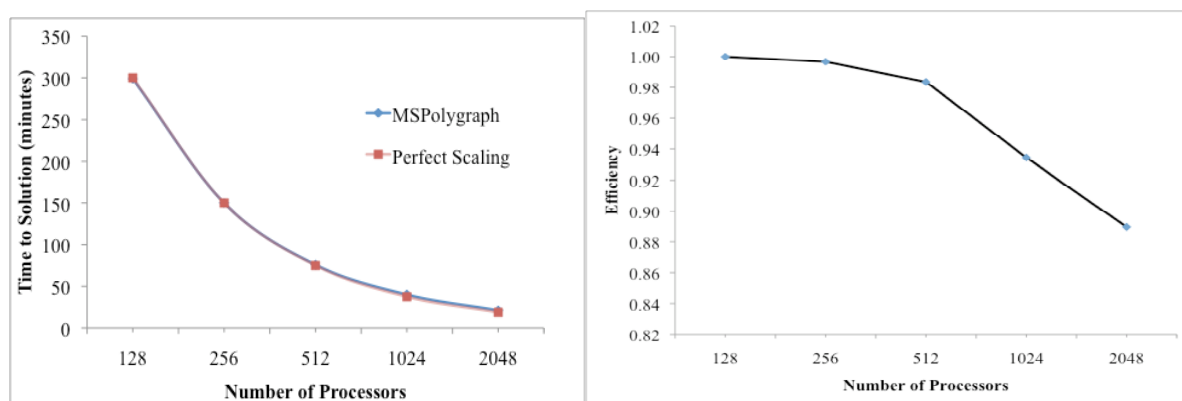


Figure 3. Time to solution (left) and parallel efficiency (right) of the *MSPolygraph* code that implements methods discussed here. The results shown are for analyzing 1,258 fully sequenced genomes against 18,929 spectra on the Chinook supercomputer at EMSL.

3. Writes data and flushes the results to its output file.
4. Issues a nonblocking receive for the next spectrum index.
5. Sends a summary message for the scored spectrum to the server.

*Scaling.* Since the processing time for a single-processor job,  $T(1)$ , takes longer to run on a single node than the job policy allowed at the time these runs, we instead generate a weak scaling curve replacing  $T(1)$  by  $128 \cdot T(128)$ . This amounts to taking the efficiency at 128 processors to be 1 for comparison purposes. We note that  $T(1)/T(128)$  must be less than or equal to 127, divvr the master processor does no work, and hence  $T(128) \geq T(1)/127$ . The scaling results are shown in Figure 2. The fall-off in efficiency shown in Figure 3 at 1,024 processors is an indicator that we are hitting inefficiency in the MPI infrastructure layer at scale, most likely due to too many messages being passed. This could also be also due to input/output bottlenecks if many processors are simultaneously writing output files. However, runtime monitoring doesn't indicate that the code is anywhere close to the I/O bandwidth limits on the machine, making it unlikely that an I/O bottleneck causes the efficiency loss.

Open source code implementing the methods discussed here for the analysis of proteomics data is available at <http://omics.pnl.gov>. The code is licensed under the Educational Community License 2.0.

### Acknowledgments

This work was supported under contracts 57271 and 54976 from the Department of Energy's Office of Advanced Scientific Computing Research (OASCR) and Office of Biological and Environmental Research (BER) to develop new approaches for computational biology in areas of national interests. The calculations were performed at the Molecular Science Computing located in the Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the Department of Energy's Office of Biological and Environmental Research and located at the Pacific Northwest National Laboratory and at the National Energy Research Scientific Computing Center at Lawrence Berkeley National Laboratory. PNNL is operated by Battelle for the U.S. Department of Energy under Contract DE-AC06-76RLO 1830.

### References

- [1] H. H. McAdams and L. Shapiro, "Circuit simulation of genetic networks," *Science*, vol. 269, pp. 650-6, Aug 4 1995.
- [2] K. L. Prather and C. H. Martin, "De novo biosynthetic pathways: rational design of microbial chemical factories," *Curr Opin Biotechnol*, vol. 19, pp. 468-74, Oct 2008.
- [3] K. A. Haynes, *et al.*, "Engineering bacteria to solve the Burnt Pancake Problem," *J Biol Eng*, vol. 2, p. 8, 2008.
- [4] D. G. Gibson, *et al.*, "Creation of a bacterial cell controlled by a chemically synthesized genome," *Science*, vol. 329, pp. 52-6, Jul 2 2010.
- [5] F. A. B. G. Bio, *et al.*, "Engineering life: building a fab for biology," *Sci Am*, vol. 294, pp. 44-51, Jun 2006.
- [6] D. A. McQuarrie, *Statistical mechanics*. New York: Harper & Row, 1976.

- [7] W. R. Cannon and M. M. Rawlins, "Physicochemical/Thermodynamic Framework for the Interpretation of Peptide Tandem Mass Spectra," *Journal of Physical Chemistry C*, vol. 114, pp. 5360-5366, Apr 1 2010.
- [8] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, pp. 379-423, 1948.
- [9] W. R. Cannon, *et al.*, "Large Improvements in MS/MS-Based Peptide Identification Rates using a Hybrid Analysis," *J Proteome Res*, Mar 30 2011.
- [10] W. R. Cannon, *et al.*, "Comparison of probability and likelihood models for peptide identification from tandem mass spectrometry data," *J Proteome Res*, vol. 4, pp. 1687-98, Sep-Oct 2005.
- [11] R. Overbeek, *et al.*, "The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes," *Nucleic Acids Res*, vol. 33, pp. 5691-702, 2005.
- [12] W. R. Cannon, *et al.*, "Evaluation of the influence of amino Acid composition on the propensity for collision-induced dissociation of model peptides using molecular dynamics simulations," *J Am Soc Mass Spectrom*, vol. 18, pp. 1625-37, Sep 2007.
- [13] A. R. Dongre, *et al.*, "Influence of peptide composition, gas-phase basicity, and chemical modification on fragmentation efficiency: Evidence for the mobile proton model," *Journal of the American Chemical Society*, vol. 118, pp. 8365-8374, SEP 4 1996.
- [14] A. Hugo, *et al.*, "Proteotyping of Microbial Communities using High Performance Optimization of Proteome-Spectra Matches," *Submitted*, 2011.
- [15] B. Latt, *et al.*, "MapReduce Implementation of a Hybrid Spectral Library- Database Search Method for Peptide Identification," *Submitted*, 2011.